Rhymes and Syntax: A Morpho-Syntactic Analysis of Czech Poetry

Silvie Cinková, Petr Plecháč, Martin Popel

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, 118 00 Praha 1, Czechia; Czech Academy of Sciences, Institute of Czech Literature of the CAS, Na Florenci 3/1420, 110 00 Praha 1, Czechia
https://orcid.org/0000-0003-4526-3915
cinkova@ufal.mff.cuni.cz

Czech Academy of Sciences, Institute of Czech Literature of the CAS, Na Florenci 3/1420, 110 00 Praha 1, Czechia
https://orcid.org/0000-0002-1003-4541
plechac@ucl.cas.cz

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, 118 00 Praha 1, Czechia
https://orcid.org/0000-0002-3628-8419
popel@ufal.mff.cuni.cz

A linguistically informed distant reading presupposes an adequate performance of Natural Language Processing tools. This article describes our evaluation of the UDPipe parser on a manually annotated sample of nineteenth-century Czech poetry in the following steps: (1) creation of a documented data set for this domain (poetry, nineteenth century, Czech); (2) domain-specific annotation decisions; (3) error analysis. The sample consisted of 29 randomly selected poems which were first automatically tagged and parsed with the UDPipe parser and then manually checked word by word. The following features were checked: word segmentation (chunking), lemmatization, part of speech assignment, assignment of more fine-grained morphological details, the position in the syntactic dependency tree (selection of the syntactic parent), as well as the label of the syntactic relation between the word and its parent. The findings were analyzed. The most typical parser errors are associated with complex noun phrases that contain other noun(s) as modifier(s), especially when these occur in a poetry-specific word order, that is, preposed to the governing noun. On the other hand, neither archaic orthography nor neologisms posed substantial issues.

Keywords: Czech poetry / distant reading / text corpora / Universal Dependencies / natural language processing / treebanks

Introduction

Some text-mining use cases benefit from reaching beyond the bag-of-words approach to extraction of lexical or grammatical patterns.[1] This is made possible by automatic morphological tagging and syntactic parsing wherever such a tool is available for the given language and achieves adequate performance within the given domain. Most parsers are run with language models that have been trained on contemporary non-fiction, and their performance is likely to decrease by the same measure that input texts deviate from those models' domains.

UDPipe, the largest Czech language model used by the best-performing Czech parser (Straka et al.), was trained on the 1990s daily Czech press (Hajič). At first glance, the main differences between this domain and that of nineteenth- and twentieth-century Czech poetry have to do with vocabulary, orthography, and word order. However, the effect of these differences on the parser performance is not predictable. The parser performance can be measured and the most typical errors can only be detected by manual annotation of a random sample and its comparison to the automatic output. While this work is time-consuming, the domain-specific annotated data could be added to the original model to increase performance on this new domain in the future—considering that this goal may turn out to require several iterations of additional annotation. In our experiments, we use the largest model, czech-pdt-ud-2.12-230717, and a smaller model based on fiction, czech-fictree-ud-2.12-230717.

Data

The data set is comprised of 29 random Czech poems from PoeTree (Plecháč et al.; Plecháč and Kolár), with a total of 6,643 tokens and 2,687 types (unique words). Most of the poems were written at the turn of the nineteenth century. About half of the represented poets belong to the Czech high-school literary canon. Most poems are rhymed. Figure 1 shows the publication dates of each poem along with its author's lifespan.

[1] The work on this article has been supported by the Czech Science Foundation grant *European Poetry: Distant Reading* 23-07727S. We have also been using data, tools, and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

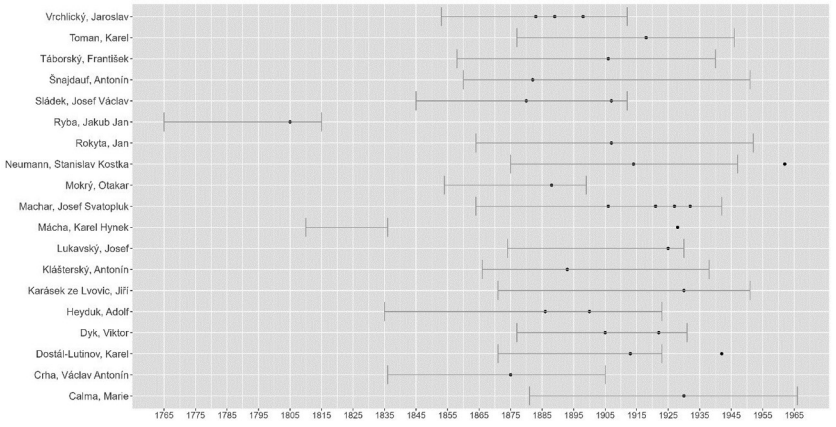

Figure 1: The PoeTree Czech sample: authors' life spans and poems' publication years.

Diachronic and stylistic language perspective

The oldest poem (1805) was written during (and in the language of) the Czech National Revival, and is therefore quite different from the later poems. Most of the nineteenth-century poems are written in somewhat modern Czech, that is, in the Czech language as it was re-established after more than a century of Germanized education and at an advanced stage of efforts to integrate the norms of a written Czech no longer in use with the spoken vernacular of the time, which was naturally perceived as low standard. The twentieth-century poems can be considered representative of (a very marked stylistic register of) truly modern Czech. The entire nineteenth century saw competing progressive as well as regressive normative trends, with the variation in poetry furthermore augmented by a rapid increase in poetic experimentation and manneristic personal style distinctions (Šlosar). Habitual modes of linguistic periodization, as a consequence, are not very helpful in the case of this poetry sample. Despite all this variation, we can still track several recurring differences between contemporary Czech prose and the language observed in this sample. This section lists a few of the resulting annotation decisions.

Spelling

Spelling variation can be found in both word stems and morphemes. In order to enable searching across different diachronic layers without altering words, we preserved the token forms while normalizing

lemmas, wherever possible, to contemporary spelling variants. For instance, rather than transcribing *nervosníma* as *nervózními*, we lemmatized using the current term *nervózní*. Whenever the modern equivalent was not instantly apparent or the word had undergone more substantial morphological changes, such as *s křeku* (*z keříku* 'from a bush') or *junoše* (*jinoch* 'lad') in the 1805 poem, we left the lemmas intact.

A prominent feature of Czech word formation—and one that presents a particularly difficult and longstanding obstacle for language processing—is compound function words. The compounding of prepositions with other parts of speech, especially nouns and adverbs, produces adverbs, particles, conjunctions, and prepositions that are written at times as discrete words and at others as prefixes, according to numerous rules with numerous exceptions (Osolsobě), thus posing challenges and spelling issues for Natural Language Processing—indeed, even for educated native speakers (Žižková). Many of these words can be found in the basic vocabulary, such as *na shledanou* 'good bye' and *zpočátku* 'initially.' Throughout the nineteenth century, little attention was paid to graphical word boundaries in general, although partial and mutually contradicting recommendations existed in grammar books. This had various consequences. For instance, the first generation of revivalists treated compound function words with complex rules depending mostly on the word formation type of the noun or adjective that followed (Dobrovský), while a later generation of grammarians tended to decompose them into discrete words (Kampelík).

The unmanageable spelling variations in the compound function words in our sample hampered lemma normalization. Whenever a compound function word consisted of two tokens, it was annotated as a syntactic relation between two words.

Punctuation and sentence splitting

The sample displays certain punctuation peculiarities: some poems combine the usual punctuation principle (syntactic segmentation) with the highlighting of rhetorical pauses (see Examples 1 and 2 below) in the manner of classic public speakers' speech notes (Pavel Kosek and Jana Pleskalová).

(1) Ale ty oči! oči smilnící!

(But those eyes! those fornicating eyes!)[2]

(2) Deset let už o tom píše, pan Vejr v Švandě Dudákovi.

(For a decade he has been writing about this, Mr. Vejr in Schwanda the Bagpiper.)

Even if the punctuation determines clause boundaries fairly well, sticking to the syntactic (and not rhetorical) principle, that is, to separate clauses (as well as conjuncts and appositions), sentence boundaries remain blurry. This applies both to poems without enjambement, where clause boundaries do not tend to cross verse boundaries (typically trochee verses, such as the extract from K. H. Mácha's *Prolog k pouti Krkonošské*, transl. *Prologue to the Riesengebirge Pilgrimage*, in Example 3 below), and to poems with long-winded clauses (often verses in prose, such as J. Karásek's ze Lvovic *Nad obrazem Marie Magdaleny* v *hradčanské Loretě*, transl. *Over the Painting of Mary Magdalene in the Hradschin Loretta*, in Example 4).

The short paratactic clauses evoke a swift narration pace and sentence boundaries do not play a role, while the syntactically long-winded clauses evoke an agitated stream of consciousness, which nevertheless unfolds within a solid syntactic scaffolding.

(3) Víc a více světnice se plní,
Hovor hlučí, kouř se z dýmek vlní;
Při stropu ho plamínku zář zlatí.

(Gradually the room is getting crowded,
The talk is loud, smoke is curling from pipes;
The glow of small flames gilds the ceiling.)

(4) V starobných ambitech, kde ztuhlá světic ctnost
V škrobených límcích španělských se vztyčuje,
Kde marně Šebestián sličný Kypící nahotu,
Drážděné šípem genitálie ukazuje,
Aby sváděl ctnostně odvrácené zraky,
Ty potměšilou vlnu ňader,
Tak měkce tajících,
Teď svůdně rozléváš,
Své tělo zase obnažuješ,

[2] All translations of poem samples are by Peter Gaffney.

Tolika tknuté milenci,
A dráždíš letí tolika ústy zlíbanou
A klínem, tolika vášněmi rozrývaným,
A očima samice, očima smilnýma.

(In the age-old ambits, where the stiff virtue of the saint rises
In starched Spanish collars,
Where in vain Sebestian, handsome,
Vibrant nudity,
His arrow-teased genitals puts on display,
To tempt the virtuously averted eyes,
You luscious wave of breasts,
So softly melting, now seductively spill,
Your body is exposed again, touched by so many lovers,
And you tease with your skin kissed by so many mouths
And a lap, torn by so many passions,
And with the eyes of a female,
With fornicating eyes.)

The strategy for annotating sentence segmentation was set to make the sentences as coherent as possible, that is, with the fewest possible standalone clauses with major ellipses.

Lexical perspective

Concerning vocabulary, some patterns of differences between PoeTree and the relevant Czech treebanks (UD_Czech-PDT and UD_Czech-FicTree, henceforth PDT and FicTree) were predictable, namely archaic word forms (*jest* 'is'; *kdys* 'long ago'), archaic words (*junoše* 'young lad'), Latin words (*Ave*; *absolvo*), and neologisms (čaroskvělý 'miraculously magnificent').

From the quantitative perspective, the overlap between case-insensitive types (unique word forms) in the PoeTree sample and the training data sets of PDT and FicTree is approximately 59 and 47 respectively, excluding proper nouns, punctuation, and symbols. That means the UDPipe parser has never seen about one half of the words that occur in PoeTree, using either model.

To allow for more qualified guesses about domain-adaptation requirements, we extracted a frequency list of all PoeTree tokens missing in PDT and a frequency list of all PoeTree tokens missing in FicTree. We compared the distributions of these types, as well as the parts of speech they belong to. In both groups, the top-ranking

PoeTree-specific tokens belong to the following parts of speech as defined by the Universal Dependencies (UD) tagging scheme: nouns, verbs, adjectives, adverbs, determiners, and pronouns. Even though the lower-ranking parts of speech ranked differently, there was no statistically significant difference between their distributions (Fisher's exact test for count data, p-value = 1).

In the next step, we extracted the symmetric difference of both lists (PoeTree-specific types that were missing either in PDT or FicTree but not in both), corresponding to 21% of their union (PoeTree-specific types missing either in PDT or FicTree or in both). From a total of 557 PoeTree-specific types, 136 were missing from PDT and 421 from FicTree. 370 of these types only occurred once in PoeTree (Figure 2). At this point we resorted to qualitative analysis.

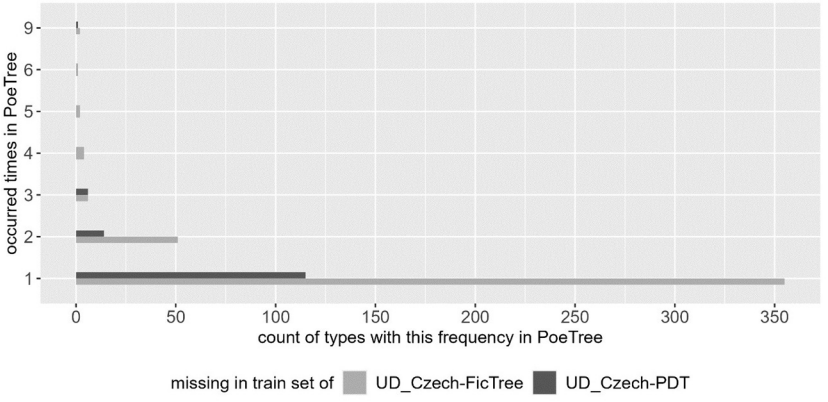

Figure 2: In which reference corpus are these PoeTree-specific words missing?

The list of missing types contained 20 types that occurred at least four times in PoeTree. Only one of them was also missing in PDT: *tobě* ('you' in the dative singular). The FicTree corpus was missing the archaic form of the third person singular of *jest* 'to be' and the vocalized form of the preposition *ku* 'towards.' Other words with minimum frequency 2 (down to Rank 87) were mostly missing in FicTree, probably because FicTree is smaller than PDT (166K tokens vs. 1M tokens in PDT). They did not seem to follow any interesting lexical or morphological pattern that would help distinguish FicTree from PoeTree. By contrast, a particularly striking pattern emerges in PoeTree compared to PDT. Here, PDT appears to have a bias against the second person singular: of 44 verb types in PoeTree that were specifically missing in PDT, six were in second-person singular form, as opposed to only two

from 143 verb types missing in FicTree. Even more strikingly, of the nine PoeTree-specific pronouns and determiners, five were second-person singular words and all were absent in PDT. They even turned out to be among the most frequent PoeTree types, which is not typical for pronouns in a pro-drop language.

Indeed, a search through the entire PDT suggested a noticeable difference in the frequency of the second person singular in PDT and in PoeTree: it detected only 45 occurrences of the singular *ty* 'you' (compared to 77 in PoeTree), nine occurrences of the singular *tvůj* 'your,' also nine in PoeTree, and 79 occurrences of the verb *být* 'to be' in the second person singular (compared to 30 in PoeTree). (It should be noted that the conjugated *to be* acts as auxiliary verb in the past and imperfective future tense.) It also detected 261 verbs in the present tense or imperative in the second person singular (101 in PoeTree), of which 95 were the fixed expression *viz* ('see,' as in 'cross reference' or 'cf.') in PDT.

Finally, we listed types missing in both PDT and FicTree. Among the most frequent types (four to six occurrences) were the archaic forms *kdys* (*kdysi* 'long ago'), *přec* (*přece* 'yet' or 'nevertheless'), *by* (*aby*, a polysemous subordinator), *chcem* (*chceme* 'we want'), and *jich* (*jejich* 'their'). The most frequent universal parts of speech (UPOS) among the hapaxes was noun, followed by verb and adjective (420, 297, and 261 occurrences respectively). Many of them were rare words or neologisms, and those belonging to common vocabulary were often either in archaic or otherwise marked forms (or second person), forming no other apparent pattern.

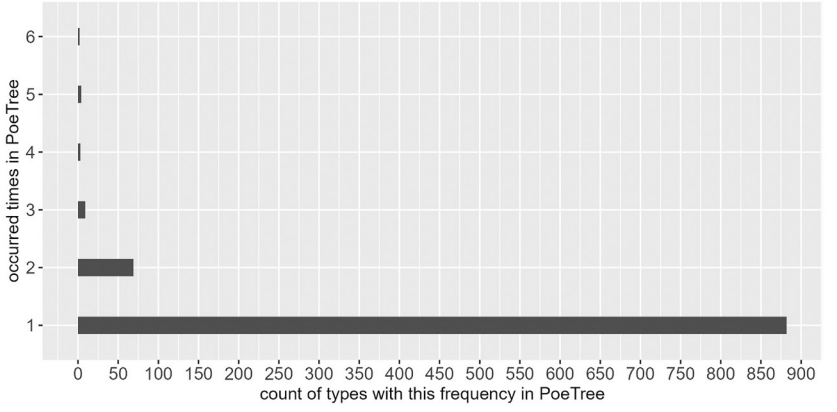

Figure 3: Distribution of PoeTree-specific words missing in both PDT and FicTree.

Syntactic perspective

While we did not make any a priori decisions concerning syntactic dependencies, we did make assumptions about how word order was likely to diverge from what is usual in modern prose or non-fiction treebanks. The observed differences are described in detail in Section 7.

The annotation process

We pre-processed the sample with the most recent version at the time of the largest Czech language model, czech-pdt-ud-2.12-220711 (Straka), employed in the UDPipe parser (Straka et al.; Straka and Straková). One annotator edited the automatic annotation node by node to come as close as possible to a manual annotation made from scratch. It was published under the title *UD_Czech-Poetry* in the Release 2.13 of the UD corpora in the LINDAT-CLARIAH repository (http://hdl.handle.net/11234/1-5287).

Evaluation results

We evaluated UDPipe's performance on the sample by comparing them to the UDPipe models based on PDT and FicTree, and then drilled into more detail using several analytical scripts in Udapi (Popel et al.). We also carried out manual error analysis.

Figure 4 presents the performance of UDPipe-PDT and UDPipe-FicTree operationalized by ten standard metrics (Kübler et al. 79–86; Zeman et al., "CoNLL"). Their values are measured as Precision (percentage of correct instances predicted by the parser), Recall (percentage of instances of gold annotation correctly predicted by the parser), and F1 (harmonic mean of Precision and Recall). They are plotted as the points of three different shapes. The first six metrics are self-explanatory, with AllTags showing the performance on morphological tagging (disregarding syntactic dependencies). The metrics UAS, LAS, MLAS, and BLEX consider each token in relation to its parent. UAS (Unlabeled Attachment Score) concentrates purely on the tree topology, which means that it only observes whether the given token is governed by the right parent. LAS (Labeled Attachment Score) considers the dependency label of the given token as well (that is, the relation to its parent). MLAS (Morphology-Aware Labeled Attachment

Score) adds UPOS and Features to the considerations. BLEX (Bilexical Dependency Score) combines content-word relations with lemmatization (but not with tags or features). The plot also shows the performance of the respective models (F1 Score) on their regular test data sets as colored bars. The performance values of both models on their regular test data sets are well above 95%. On PoeTree, the performance is generally worse, by the largest margin in Sentences and MLAS.

UDPipe-PDT and UDPipe-FicTree perform very similarly on PoeTree, with UDPipe-PDT scoring slightly better than UDPipe-FicTree in general and even substantially better in Sentences, UFeats, and MLAS. Therefore, UDPipe-PDT appears to be the parser of first choice for PoeTree and we limit manual error analysis in the next sections to the output of UDPipe-PDT.

Error analysis

Figure 4 reveals that the lowest scoring metric is Sentences, that is, the recognition of sentence boundaries. This is indeed neither surprising, given the a priori observations of punctuation and sentence splitting, nor extremely interesting, since sentence boundaries in poetry are often disputable even to a human annotator. For further error analysis, we have therefore aligned the manual and automatic word-to-word and re-segmented the automatic annotation to matching chunks of text. In this setup, we counted and classified the mismatches between manual and automatic annotation.

The most frequent error is the choice of parent (546, that is, ca. 8% of tokens), of which 395 are not combined with any labeling error. This also corresponds to Figure 4, where the second lowest scoring metric is MLAS, the combination of tree topology (choice of parent) and the syntactic and morphological labels in the given token. It also confirms that topological errors are not to blame on sentence splitting alone.

Of the 50 most frequent errors listed by the official UD evaluation script (Straka and Popel), 26 are dependency-labeling (deprel) errors, 13 are tokenization errors in thus far unseen contracted forms with unstable orthography, 11 are feature-labeling errors (Ufeats), four are part-of-speech errors, and four are lemma errors.

The lemma errors revolve around the so-called canonic number for the base form in pronouns (e.g., *náš* ‘our’ as *náš* ‘our’ vs. *můj* ‘my’) and reveal the general need for permanent data harmonization against the ultimate morphological lexicon (Hajič et al., “MorfFlex”).

The most frequent UPOS error (24 occurrences, or 0.4%) concerned the blurry distinction between adverbs and particles (also suggesting inconsistencies in the manual annotation of different corpora), and the similarly blurry distinction between coordinating conjunctions used within a single sentence compared to those used across sentence boundaries (to be marked as sentential adverbs).

We also found that the most frequent features errors were not really errors but innovations encouraged by the Czech UD coordinators: to date, homonymous word forms have not been disambiguated in the PDT and FicTree data sets (e.g., Gender=Fem,Neut), unlike the PoeTree sample (Gender=Fem or Gender=Neut). These two approaches differ by the extent of contextualization. While the earlier approach deliberately relied on as little context as possible, the more recent developments in machine learning are likely to master context-based morphological disambiguation even across sentence boundaries. A prominent example of this change might be the disambiguation of active verb participles (used to form past tense): the neutral plural is homonymous with feminine singular, and Czech is a pro-drop language, which means that the coreferential antecedent of the dropped subject must often be tracked back across sentence boundaries.

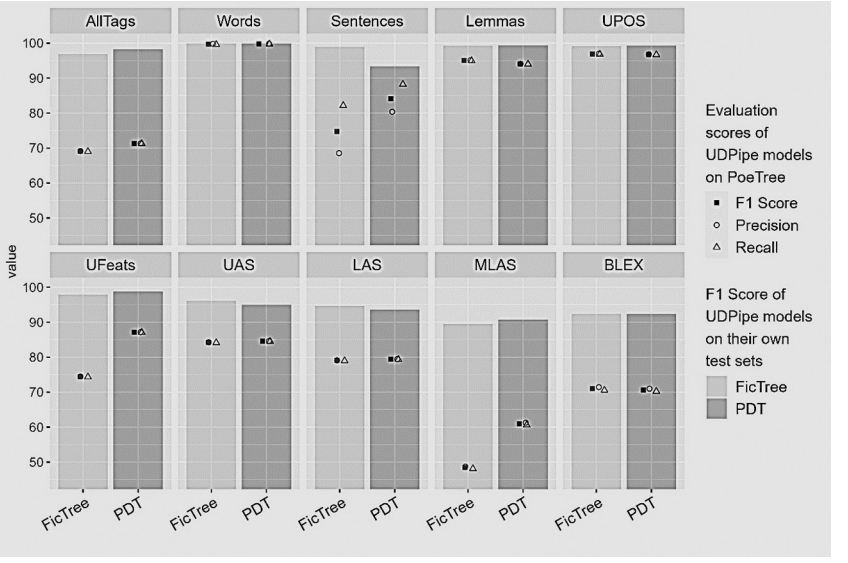

Figure 4: Model evaluation.

Since the aforementioned errors are not entirely errors, occur only rarely, or can be automatically corrected in the model training data, after which they are likely to present themselves correctly, what remains is tree topology and dependency labeling (syntactic parsing). Focusing on dependencies also makes sense, given that syntactic dependencies represent one of the advantages of extracting information from treebanks with comparison to carrying out linear searches. In the context of information-extraction use case, we find it appropriate to emphasize errors in phenomena that are likely to hamper the extraction of relevant patterns (such as convoluted attributive structures) over errors that may be frequent but do not necessarily affect rule-based extraction of noun modifiers or predicates and their arguments and adjuncts. Such largely irrelevant errors may involve punctuation, coordination vs. parataxis mismatches, or inconsistent labeling of prepositional noun modifiers, such as we find with nmod (noun modifier) vs. obl (oblique case).

Most prominent parsing errors

Labeling confusion as weighted centrality degree in a network of labels

The LAS results are best explained as a directed network graph (Figure 5) of dependency labels (deprels), with emphasis on their weighted degree centrality. Each edge connects a source node (human-assigned deprel) with a target node (deprel automatically assigned by UDPipe) on the same token, with the frequency of the given deprel combination in the same source-target direction increasing the edge weight. The number of outgoing edges along with their weights constitutes the weighted out-degree centrality of each deprel.

In this scheme, deprels can have out-degree centrality only when used in the gold annotation, whereas they only have in-degree centrality when they are used in the automatic parsing. Hence, the top-ranking deprels listed in Table 1 and highlighted in Figure 5 are gold-annotation deprels that UDPipe labeled with the wrong deprel, in addition to attaching them to the wrong parent. Each node in this graph represents one syntactic label. The nodes are connected with directed edges (arrows). Each arrow starts in the gold annotation and points to its mismatched label in the automatic annotation, respectively. Dotted edges connect the top 20% of gold annotation with the most frequent mismatches (totaled across all mismatched labels).

The thick highlighted source-target edges in Figure 5 convey which deprels are frequently confused in both directions, such as, for instance, nmod and obl (ranking 1 and 2 in Table 1). Both denote a noun or noun group, possibly even introduced by a preposition. This modifier is called nmod (noun modifier) when modifying a noun, such as *John* in *letter to John*, but obl (oblique case) when modifying a verb, such as in *write to John*. In a vague context such as *write a letter to John*, the modifier *John* can be attached to either, while in *give the letter from/by Mary to John*, we would rather attach *Mary* to the noun *letter* as nmod than to the verb *give* as obl.

It does not come as a surprise that conj (conjunction) and root are strongly interconnected in both directions: in complex sentences with several clauses, the parser easily fails to identify the main predicate. Quite symptomatically, root is also connected with advcl (adverbial clause, subordinate clause), parataxis (coordination of two main clauses without a conjunction), and orphan (clause with an elided predicate). The strong associations of conj with obl, nsubj (nominal subject), and, to a lesser extent, obj (direct object), indicate misrecognized coordinations of nouns.

The second strongest association with root is nsubj, which can be easily accounted for by the fact that the UD scheme prefers content words as parents of function words (e.g., nouns govern prepositions), while at the same time regarding copula verbs as auxiliary words. In copula predicates, therefore, the root is the predicate noun (Figure 6), which may be confused in turn with the subject (nsubj).

Ultimately, the strongest confusion emerges between the aforementioned nmod and obl. Since our statistic considers only labeling mismatches on incorrectly attached nodes, we can generally assume that nmod cases in the automatic sample are governed by nouns (since the parser has learned that nmod only modifies nouns) and obl cases are governed by verbs. This implies that a fraction of nouns in attributive positions or positions of verb arguments or adjuncts will be systematically lost when searching the poetry data with syntactic corpus queries. It is worthwhile to investigate whether this confusion occurs randomly or according to a pattern.

deprel	weighted out-degree centrality	parent UPOS NOUN	parent UPOS VERB
nmod	68	+	-
obl	47	-	+
root	43	-	-
nsubj	32	-	+
conj	31	+	+
advcl	29	-	+
amod	28	+	-

Table 1: Highest weighted out-degree centrality. The last two columns illustrate the possible distribution of the deprels among nouns and verbs as parent tokens.

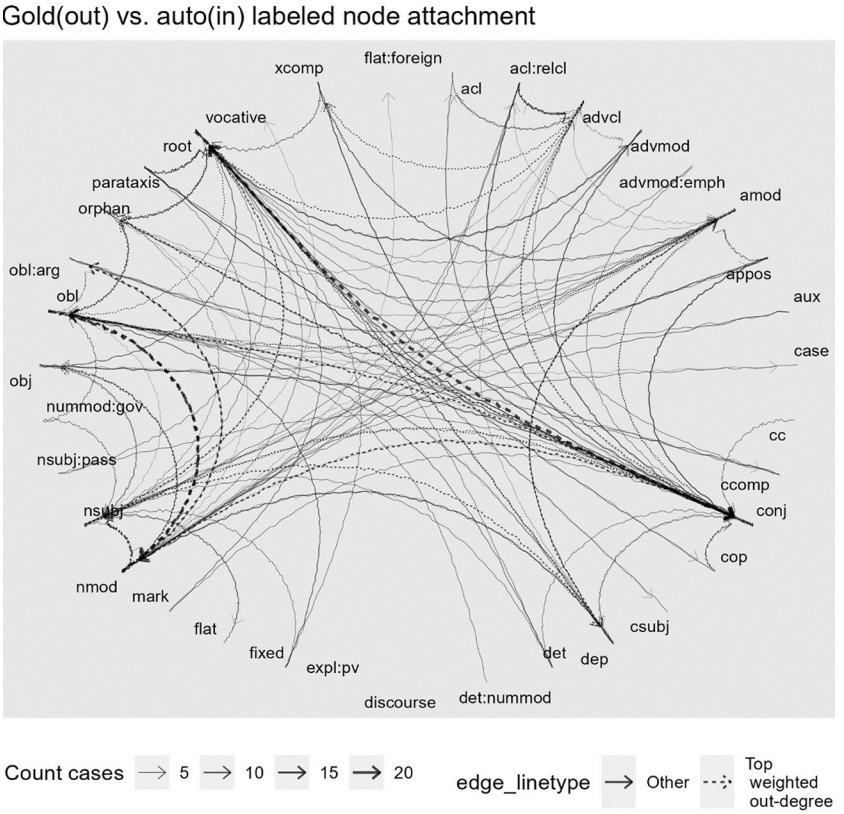

Figure 5: The most prominent labeling errors in a network graph of tokens with the wrong parent as well as the wrong dependency label.

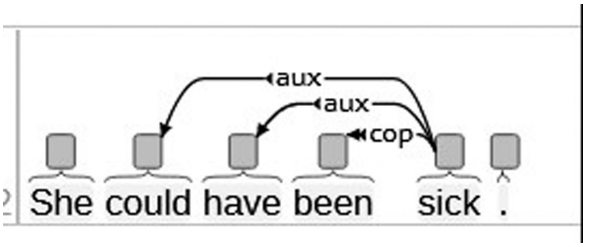

Figure 6: Copula predicate. The predicate noun and hence the sentence root is *sick*, while *she* is the subject (Marneffe et al.).

Labeling errors on nouns in attributive position

The high weighted out-degree centrality of nmod means that nmod UDPipe kept assigning other labels to nodes that should have been nmod. It hence makes sense to examine errors from the perspective of gold data, that is, to concentrate on nouns and their noun attributes.

When concentrating on nouns and their attributes, we get the following picture: the PoeTree sample contains 501 cases of attributive nouns (i.e., nmod). Of those, 169 (34%) were attached to an incorrect parent. Of 117 attributive nouns in a prepositional case, 49 (41%) were misrecognized. Of 384 attributive nouns in a direct case, 120 (31%) were misrecognized. However, when the case was genitive and the attribute noun was preposed, as many as 51 of 52 cases (98%) were misrecognized.

Comparing that with adjectival attributes, we observed only 117 of 877 incorrect parent attachments (13%). When the adjective preceded the noun, 63 of 575 attributes (11%) were incorrectly attached; when it followed the noun, the number was 54 of 302 (17%). Parser performance decreased in proportion to the distance between tokens. However, the data was too sparse to be statistically significant (for a token distance of 3 or more, the number was 19 of 25 errors if postposed, and 6 of 36 if preposed).

Preposed genitive attributes

The analysis has shown that the parser failed most dramatically with preposed genitive attributes, apparently because it had never spotted them in the training data.

In current Czech prose, it is not uncommon for noun attributes to take the form of another noun in the genitive case. The noun in genitive often denotes the agent or patient of an event (*hledání odpovědi* 'the search for an answer'), the owner or bearer (*planeta opic* 'planet of the apes'), or a quantified mass or set (*pytel brambor* 'sack of potatoes'). Nevertheless, in all these cases, the genitive follows the head noun. In the entire PDT, there are only two cases of a preposed genitive attribute. One is the lexicalized expression *svého druhu* 'of sorts'; the other concerns attributive nouns modified by a cardinal numeral, which in Czech requires the genitive of the governing noun (Figure 7). In this last case, one could argue that word order is slightly marked, emphasizing the amount, whereas in the unmarked order the genitive noun follows the head noun (see Section 8.1).

In poetry, on the other hand, the preposed genitive attribute is a legitimate structure, given its 10% proportion of all attributive nouns in our sample. This alone–its 98% error rate–calls for a domain adaptation of the language model to poetry.

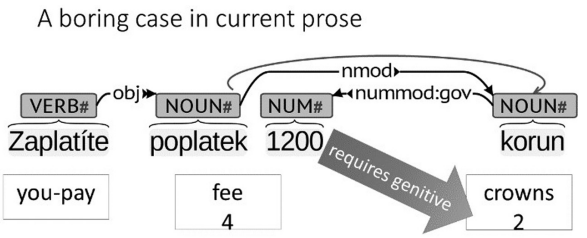

Figure 7: Preposed noun genitive in current Czech. The numbers 4 and 2 mean accusative and genitive.

Even with current Czech, the parser gets confused (Figure 7), regarding both *poplatek* 'fee' and *korun* 'crowns' as verb arguments, rather than two direct objects (obj), which the annotation scheme does not allow (a verb clause can only have one instance of a subject and object). Errors of this kind appear systematically in the poetry data (Figure 8).

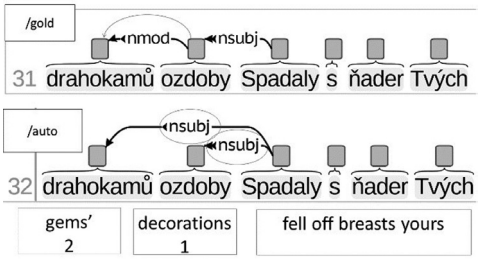

Figure 8: Preposed genitive attribute in poetry.

As Figures 9, 10, and 11 show, virtually any clause chunk can land between the preposed genitive attribute and head noun, resulting in additional parsing errors in the vicinity. This is why UDPipe again mistakes the genitive attribute, in Figure 10, for a second subject apart from its head noun (the true subject). In Figure 9, UDPipe (bottom line) has not recognized any syntactic dependency relation between the genitive (*Madonna's*) and head noun (*face*). The same applies to Figure 11, where it has missed the relation between *hair* (in the genitive) and *flood*.

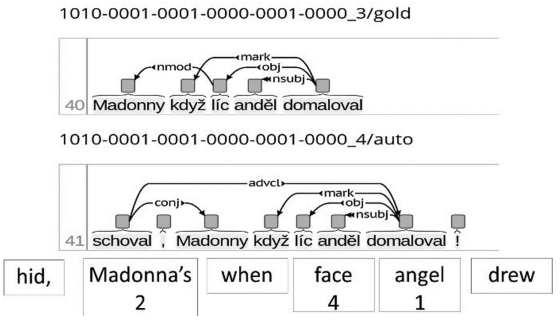

Figure 9: Discontinuous attributive sequence.

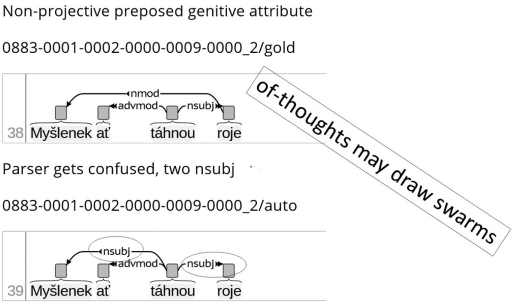

Figure 10: A whole clause between attributive genitive and its head noun.

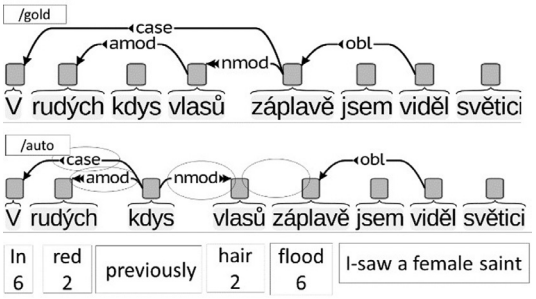

Figure 11: Very disrupted parsing.

Comparison of parsing errors in PoeTree and PDT

Qualitative findings from the previous section suggest that the specific constraints on prosody and meter may require poetic texts to allow longer distances between tokens and their modifiers (edge lengths, measured in tokens), as well as specific word order patterns. This section investigates the distance between several frequent syntactic dependencies, the order of their members, and the performance of the UDPipe-PDT model on two data sets: the PoeTree sample and PDT-test set (on which the performance of UDPipe-PDT was measured).

Performance on preposed genitive attributes

As Figure 12 shows, most preposed genitive attributes occur immediately before the head noun, within the maximum distance (edge length) -6 in PDT-test and -5 in PoeTree. The red bars in the blue-red pairs are lower than the blue ones in both PDT-test and PoeTree, but the difference is smaller in PDT than in PoeTree, which implies higher recall in PDT than in PoeTree. At the same time, orphaned red bars occur to the right of zero in both data sets. These are precision errors, and they are markedly fewer in the PDT-test sample.

By and large, the distributions of edge lengths are almost identical. At this point we should note that in PDT, unlike PoeTree, the preposed genitives are the product of grammatical congruence with a genitive-requiring cardinal numeral (denoting containers, substances, currencies, or metric units, see Section 7.3). Therefore it comes as no surprise that UDPipe processes them much better in PDT than in PoeTree. Also disregarding the lexical patterns and looking at raw

frequencies, the preposed genitives are clearly overrepresented in poetic texts compared to PDT, given that the PDT-test set is almost 27 times larger than the PoeTree sample and its occurrence of preposed genitives is only approximately double.

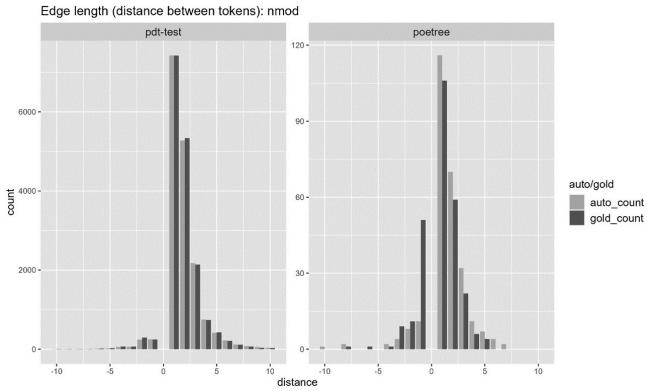

Figure 12: UDPipe-PDT's performance on the preposed genitive attribute in PoeTree and PDT-test.

Performance on any noun attribute

Generally speaking, noun attributes preceding nouns are overrepresented in poetic texts, and UDPipe has a precision issue with postposed noun attributes in PoeTree and a recall issue with preposed noun attributes (slightly above half of the approximately 50 items immediately before head nouns would be genitives).

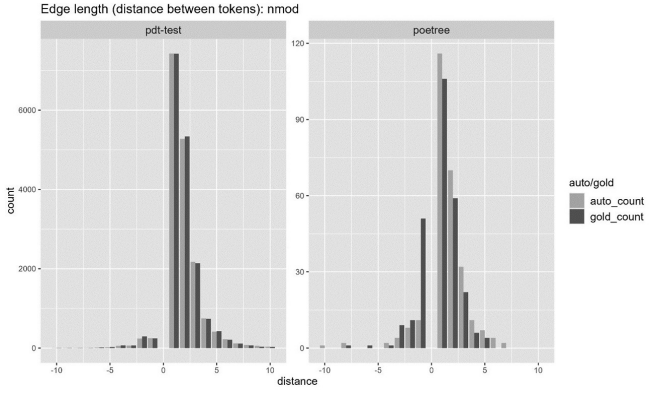

Figure 13: UDPipe-PDT's performance on noun attributes in PoeTree and PDT-test.

Performance on adjective attributes

The adjective attributes apparently are relatively more frequent in PoeTree than in PDT, but UDPipe-PDT processes them well, although the overall performance of UDPipe-PDT on PoeTree is slightly lower than on PDT-test, with errors both in precision and in recall, and both left and right from the governing noun.

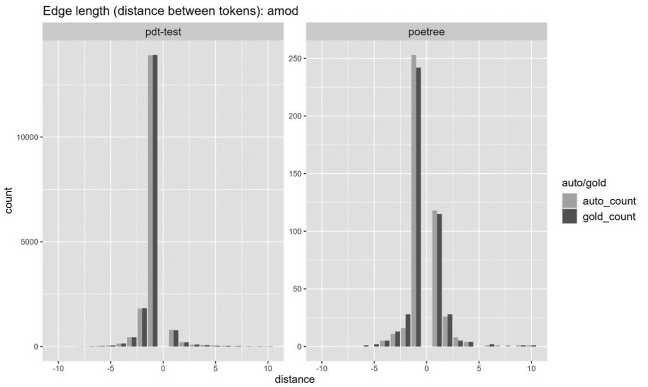

Figure 14: UDPipe-PDT's performance on adjective attributes in PoeTree and PDT-test.

Performance on clause subjects

The distribution of edge lengths for the clause subject is apparently identical in both data sets. On PDT-test, UDPipe-PDT tends to produce precision errors, while on PoeTree both error types occur. Interestingly, the overall performance appears slightly higher on PoeTree.

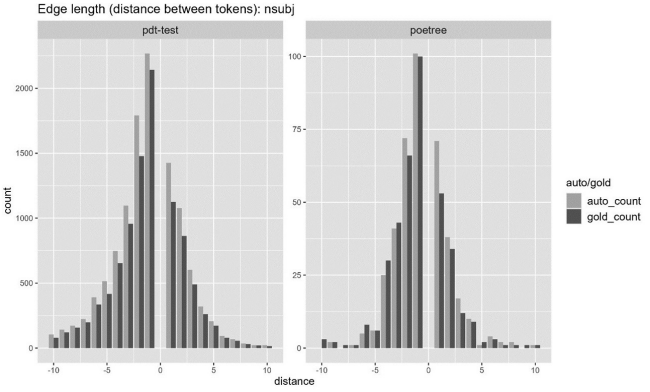

Figure 15: UDPipe-PDT's performance on subjects in PoeTree and PDT-test.

Performance on direct objects

Direct objects occur apparently more often immediately before their governing verb in PoeTree than they do in PDT-test. UDPipe-PDT performs slightly worse on PoeTree than on PDT-test, in all positions, but the difference is not dramatic.

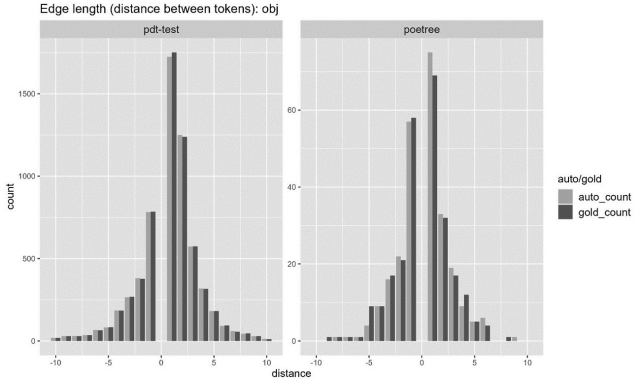

Figure 16: UDPipe-PDT's performance on direct objects.

Performance on prepositional objects and adverbials

Distributions are similar for prepositional objects and adverbials, with one interesting observation: objects immediately following the verb are rather rare, especially in PDT-test, and UDPipe-PDT has a severe precision problem (too many false positives) on both datasets, across positions.

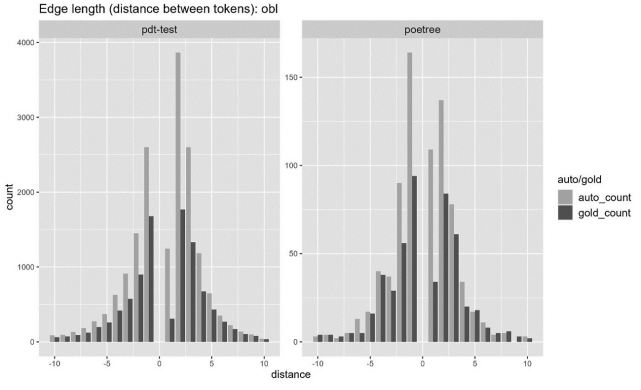

Figure 17: UDPipe-PDT's performance on nouns with prepositions modifying verbs.

Discussion and conclusion

We have evaluated the performance of the UDPipe parser with the largest Czech model based on the Prague Dependency Treebank (Hajič et al., "Prague") converted to Universal Dependencies (Zeman et al., "Universal"), and performed a semi-manual error analysis focused on parts of speech and dependency relations that are most likely to occur in corpus queries to extract information from texts in text-mining or distant reading research tasks.

Czech poetry makes ample use of the free word order that is a feature of the Czech language. Hence, PoeTree contains structures that do not normally occur, and UDPipe-PDT fails to parse them correctly because it has never spotted them in the training data. These structures are not random but recurrent, and therefore it is important to, first, identify and tackle them as parsing issues, and second, provide manually annotated data to the UDPipe model training pipeline to improve UDPipe's performance on poetry.

WORKS CITED

Dobrovský, Josef. *Ausführliches Lehrgebäude der Böhmischen Sprache, zur gründlichen Erlernung derselben für Deutsche, zur vollkommenern Kenntniß* für Böhmen. Prague, Johann Herrl, 1809.

Hajič, Jan. "Complex Corpus Annotation: The Prague Dependency Treebank." Jazykovedný ústav L. Štúra, *SAV*, 2004, https://ufal.mff.cuni.cz/pdt2.0/publications/Hajic2004.pdf. Accessed 24 Jan. 2024.

Hajič, Jan, et al. "MorfFlex CZ 2.0." *LINDAT/CLARIAH-CZ*, 2020, http://hdl.handle.net/11234/1-3186. Accessed 24 Jan. 2024.

Hajič, Jan, et al. "Prague Dependency Treebank 2.0." *Linguistic Data Consortium*, 2006, https://ufal.mff.cuni.cz/pdt2.0/. Accessed 24 Jan. 2024.

Kampelík, František Cyril. *Čechoslovan, čili národní jazyk v Čechách, Na Moravě, ve Slezku a Slovensku*. Prague, Jan Hostivít Pospíšil, 1842.

Kübler, Sandra, et al. *Dependency Parsing*. Springer, 2009.

Marneffe, Marie-Catherine de, et al. "Syntax: General Principles–The Status of Function Words." *Universal Dependencies Guidelines*, 2017, https://universaldependencies.org/u/overview/syntax.html#the-status-of-function-words. Accessed 24 Jan. 2024.

Osolsobě, Klára. *Česká morfologie a korpusy*. Prague, Karolinum, 2014.

Kosek, Pavel, and Jana Pleskalová. "Spřežkový Pravopis." *CzechEncy–Nový encyklopedický slovník češtiny*, edited by Petr Karlík et al., Brno, Masarykova univerzita, 2017, https://www.czechency.org/slovnik/SPŘEŽKOVÝ PRAVOPIS. Accessed 24 Jan. 2024.

Plecháč, Petr, and Robert Kolár. "The Corpus of Czech Verse." *Studia Metrica et Poetica*, vol. 2, no. 1, 2015, pp. 107–118, https://doi.org/10.12697/smp.2015.2.1.05. Accessed 24 Jan. 2024.

Plecháč, Petr, et al. *PoeTree. Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian and Spanish. 0.0.1*. Zenodo, 2023., https://zenodo.org/records/10008459. Accessed 24 Jan. 2024.

Popel, Martin, et al. "Udapi: Universal API for Universal Dependencies." *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*, edited by Marie-Catherine de Marneffe et al., Northern European Association for Language Technology, 2017, pp. 96–101.

Straka, Milan. "Universal Dependencies 2.12 Models for UDPipe 2." *LINDAT/CLARIAH-CZ*, 2023, http://hdl.handle.net/11234/1-5200. Accessed 24 Jan. 2024.

Straka, Milan, and Martin Popel. "Eval.Py. 1.2." *GitHub*, 2023, https://github.com/UniversalDependencies/tools/blob/master/eval.py. Accessed 24 Jan. 2024.

Straka, Milan, and Jana Straková. "UDPipe 2." *LINDAT/CLARIAH-CZ*, 2022, http://hdl.handle.net/11234/1-4816. Accessed 24 Jan. 2024.

Straka, Milan, et al. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing." *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari et al., European Language Resources Association, Paris, 2016, pp. 4290–4297, https://aclanthology.org/L16-1680. Accessed 24 Jan. 2024.

Zeman, Daniel, et al. "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies." *Proceedings of the CoNLL 2018 Shared Task*, edited by Daniel Zeman and Jan Hajič, Kerrville (TX), The Association for Computational Linguistics, 2018, pp. 1–21, http://www.aclweb.org/anthology/K18-2001. Accessed 24 Jan. 2024.

Zeman, Daniel, et al. "Universal Dependencies 2.12." *LINDAT/CLARIAH-CZ*, 2023, http://hdl.handle.net/11234/1-5150. Accessed 24 Jan. 2024.

Žižková, Hana. "Compound Adverbs as an Issue in Machine Analysis of Czech Language." *Journal of Linguistics / Jazykoedný* Časopis, vol. 68, no. 2, 2017, pp. 396–403, https://doi.org/10.1515/jazcas-2017-0049. Accessed 24 Jan. 2024.

Rime in skladnja: oblikoskladenjska analiza češke poezije

Ključne besede: češka poezija / oddaljeno branje / besedilni korpusi / Universal Dependencies / obdelava naravnega jezika / odvisnostne drevesnice

Oddaljeno branje, ki upošteva jezikoslovna spoznanja, predpostavlja ustrezno delovanje orodij za obdelavo naravnega jezika. Članek prikaže evalvacijo razčlenjevalnika UDPipe na primeru ročno označenega vzorca češke poezije 19. stoletja v naslednjih korakih: (1) ustvarjanje dokumentiranega nabora podatkov za to področje (poezija, 19. stoletje, češčina); (2) odločitve o označevanju,

specifične za področje; (3) analiza napak. Vzorec je obsegal 29 naključno izbranih pesmi, ki so bile najprej samodejno označene in razčlenjene z razčlenjevalnikom UDPipe, nato pa so bile oznake ročno preverjene za vsako posamično besedo. Preverjene so bile naslednje značilnosti: segmentacija besed (razdelitev), lematizacija, dodelitev oblikoskladenjskih oznak, dodelitev natančnejših morfoloških oznak, dodelitev položaja v skladenjskem drevesu (izbor nadrejenega elementa) in oznaka skladenjskega razmerja med besedo in njenim nadrejenim elementom. Ugotovitve smo analizirali; najpogostejše napake razčlenjevalnika so povezane s kompleksnimi samostalniškimi besednimi zvezami, ki vsebujejo druge samostalnike kot modifikatorje, še posebej, če se ti pojavijo v besednem redu, specifičnem za poezijo, npr. kot določilo samostalniškega jedra. Po drugi strani niti arhaični pravopis niti neologizmi niso predstavljali bistvenih težav.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.162.3.09-1"18":004
DOI: https://doi.org/10.3986/pkn.v47.i2.04